Choice of generalized linear mixed models using predictive crossvalidation

Braun, Julia ; Held, Leonhard ; Sabanés Bové, Daniel

Abstract: The choice of generalized linear mixed models is difficult, because it involves the selection of both fixed and random effects. Classical criteria like Akaike's information criterion (AIC) are often not suitable for the latter task, and others which are useful in linear mixed models are difficult to extend to the generalized case, especially for overdispersed data. A predictive leave-one-out crossvalidation approach is suggested that can be applied for choosing both fixed and random effects, even in models with overdispersion, and is based on proper scoring rules. An attractive feature of this approach is the fact that the model has to be fitted just once to the data set, which makes computations fast and convenient. As the calculation of the leave-one-out predictive distribution is not possible analytically, it is shown how an iteratively weighted least squares algorithm combined with some analytic approximations can be used for this task. A simulation study and two applications of the methodology to binary and count data are provided, as well as comparisons with two other methods.

DOI: https://doi.org/10.1016/j.csda.2014.02.008

Choice of generalized linear mixed models using predictive crossvalidation[☆]

Julia Braun[*], Daniel Sabanés Bové, Leonhard Held

Division of Biostatistics, Institute for Social and Preventive Medicine,University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

Abstract

The choice of generalized linear mixed models is difficult, because it involves the selection of both fixed and random effects. Classical criteria like Akaike's information criterion (AIC) are often not suitable for the latter task, and others which are useful in linear mixed models are difficult to extend to the generalized case, especially for overdispersed data. A predictive leave-one-out crossvalidation approach is suggested that can be applied for choosing both fixed and random effects, even in models with overdispersion, and is based on proper scoring rules. An attractive feature of this approach is the fact that the model has to be fitted just once to the data set, which makes computations fast and convenient. As the calculation of the leave-one-out predictive distribution is not possible analytically, it is shown how an iteratively weighted least squares algorithm combined with some analytic approximations can be used for this task. A simulation study and two applications of the methodology to binary and count data are provided, as well as comparisons with two other methods.

Keywords: Predictive model choice, proper scoring rules, Poisson regression, logistic regression, conditional AIC, overdispersion.

[☆]Data and functions used in this paper can be found as online supplementary material.

[*]Corresponding author. Tel.: +41 44 634 46 27

Email addresses: julia.braun@ifspm.uzh.ch (Julia Braun), daniel.sabanesbove@ifspm.uzh.ch (Daniel Sabanés Bové), leonhard.held@ifspm.uzh.ch (Leonhard Held)

1. Introduction

Model choice in linear or generalized linear models is a relatively straightforward task, and various criteria and techniques are available. If, however, these models are extended to contain random effects to accommodate e.g. longitudinal data, choosing a model becomes much more challenging. One reason for this is the fact that in addition to the selection of covariates, a decision on the kind and number of random effects has to be made. Classical criteria like Akaike's information criterion (AIC, [1]) or the Bayesian information criterion (BIC, [35], [7]) are not sufficient for this task and must be adapted.

Before applying any criterion for model choice, a decision on the focus of the desired analysis has to be made, namely if the main interest lies on the fixed effects (population level) or if information on the random effects (individual or cluster level) is desired. In the case of linear mixed models, the choice of fixed effects using a BIC version suitable for unbalanced longitudinal data is suggested ([31]). For choosing random effects, a boundary Laplace approximation to obtain a BIC version including an additional term for boundary correction can be used ([32]).

In terms of the AIC, the consequences of the main focus of inference are highlighted, showing that the generally known, classical version of the AIC – the marginal AIC – should only be applied for the selection of fixed effects ([38]). For deciding on the inclusion of random effects, the conditional AIC (cAIC) is introduced, for which the effective degrees of freedom needed in the penalty term can be calculated ([23]). This concept is extended to deliver more reliable results ([26]), but the numerical calculation is quite involved and time-consuming ([20]). Details on likelihood ratio tests for linear mixed models are also given ([9]).

Unfortunately, all these concepts only relate to linear mixed models and are difficult to extend to generalized linear mixed models. A Bayesian approach to the simultaneous selection of fixed and random effects via zero-inflated (truncated) normal priors on fixed effects and on elements of the decomposed random

effects covariance matrix is presented ([6]). Fence methods for the selection of the fixed effects in generalized linear mixed models are suggested ([24], [30]). These methods choose models from a range of candidate models by setting and subsequently restricting boundaries of some suitable criterion ([7, p. 273]). An analytic deduction of the cAIC is impossible ([13]), but the authors suggest an asymptotic approximation which includes the effective degrees of freedom ([27]). They note, however, that their asymptotic approximation may not be reliable in certain settings and propose using bootstrap methods instead. An asymptotically unbiased estimator of the cAIC for use with generalized linear mixed models is presented ([42]), which seems to be quite similar to the above mentioned approximation ([13]). Another unbiased estimator of the cAIC to be used for Poisson regression models is proposed, which involves a high number of model fits and might thus be unsuitable for large data sets ([25]) .

In this article, we introduce an alternative approach to selecting generalized linear mixed models for longitudinal data from a predictive point of view. By using mean crossvalidated proper scores ([18]) as criterion for model choice, both fixed and random effects can be selected leading to a model with the best predictive abilities. The crossvalidated logarithmic score is closely related to the AIC in linear models ([36], [33]) and to the cAIC in linear mixed models ([4]), so that its application in the case of generalized linear mixed models seems promising.

Other than in the linear mixed model, the (leave-one-out) predictive distribution that is necessary for the calculation of the proper scores cannot be deducted analytically. To solve this problem, we propose to use an iteratively weighted least squares (IWLS) algorithm with prior distribution ([16]), complemented by some analytic approximations. To shorten the computation time, we reduce the number of necessary model fits to just one, using "mixed" crossvalidation ([28]). This approach is by far less time-consuming than full leave-one-out crossvalidation, and it has been shown empirically that the results from both crossvalidation approaches are comparable for the linear mixed model ([4]).

This article is organized as follows: We review the basics of generalized lin-

ear mixed models in Section 2 and show the proper scoring rules needed for comparing predictive distributions in Section 3. The predictive crossvalidation approach based on a Bayesian IWLS algorithm is presented and outlined specifically for logistic regression and Poisson regression (with and without overdispersion) in Section 4. Results from a simulation study to investigate the method's properties are shown in Section 5. Applications to binary and count data are discussed in Section 6, followed by a comparison with two other approximate estimators of the cAIC and with a full leave-one-out crossvalidation. Section 7 adds a summary and some general discussion.

2. Generalized linear mixed models

Generalized linear mixed models for longitudinal data are generally defined as follows (see for example [15]): Assume that each individual $i = 1, \dots, I$ provides observations y_{ij} at time points t_j with $j = 1, \dots, J$. For simplicity, we assume that the time points are the same for each individual, but this is not a necessary precondition. Let the vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain covariates relating to the fixed and random effects, respectively, then the linear predictor is defined as

$$\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i,$$

with p fixed effects $\boldsymbol{\beta}$ and q random effects $\boldsymbol{b}_i$. The conditional expected value $\mu_{ij} = \mathrm{E}(y_{ij} \,|\, \boldsymbol{b}_i)$ is related to the linear predictor via an appropriate link function g, so that $g(\mu_{ij}) = \eta_{ij}$.

The two non-Gaussian regression models that are applied most often in this context are binary logistic and log-linear Poisson regression models. In the case of logistic regression, each observation y_{ij} has a Bernoulli distribution with probability

$$p_{ij} = P\{y_{ij} = 1\} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}. \tag{1}$$

The log-linear Poisson model assumes the expectation

$$\lambda_{ij} = \exp(\eta_{ij}).$$

Overdispersion can be included in a model by estimating an additional random effect for each observation ([8, p. 293]). The mixed Poisson model with overdispersion has a linear predictor of the form

$$\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + a_{ij},$$

where $a_{ij} \sim N(0, \xi^2)$ represents the additional random effect for each observation y_{ij}, and $\boldsymbol{b}_i$ and a_{ij} are mutually independent. Information on fitting and interpreting generalized linear mixed models as well as other suitable models for discrete longitudinal data can be found in the literature, e.g. [29].

3. Proper scoring rules

As stated in the introduction, the selection of random effects in generalized linear mixed models based on any selection criterion that is currently available is difficult. Therefore, we suggest to use proper scoring rules as a criterion for model choice. They are a simple, yet effective instrument for assessing predicive distributions, allowing the choice of both fixed and random effects. In this paper, all scores are positively oriented, so that a larger score denotes a model with better predictive abilities. By taking into account not only the distance between a point prediction and the true value, but also the predictive variance, proper scores simultaneously cover both sharpness, i.e. the concentration of a predictive distribution, and calibration, i.e. the consistency between the predictive distribution and the actual observations ([41]). General information on the concept of proper scoring rules as well as examples for the case of continuous predictive distributions can be found ([18], [17]).

A well-known proper scoring rule for binary predictions is the Brier score ([5]). It is also called the quadratic score and is defined as

$$\mathrm{BS}(Y, y_{\mathrm{obs}}) = -(p_Y - y_{\mathrm{obs}})^2,$$

where p_Y stands for the predicted probabiliy of the outcome and $y_{\mathrm{obs}} \in \{0, 1\}$ is the actual observation.

The proper logarithmic score (LS) is well suited to assess the predictive abilities of any regression model if the density f_Y of the predictive distribution Y is known. It is defined as the value of the log density of Y at the actually observed value y_{obs}:

$$\text{LS}(Y, y_{\text{obs}}) = \log f_Y(y_{\text{obs}}). \tag{2}$$

An alternative to the LS is the Dawid-Sebastiani score (DSS) ([11]) which is used as predictive model choice criterion ([21]). It is defined as

$$\text{DSS}(Y, y_{\text{obs}}) = -\frac{1}{2}\left\{\log(\sigma_Y^2) + \left(\frac{y_{\text{obs}} - \mu_Y}{\sigma_Y}\right)^2\right\} \tag{3}$$

and has the advantage that only the first two central moments μ_Y and σ_Y^2 of the predictive distribution of Y are necessary for its calculation. Alternative model assessment tools for use with models for count data are given ([10]).

4. Predictive crossvalidation

Conducting a full crossvalidation often turns out to be very time-consuming and in some cases even impossible due to the size of the respective data set and the complexity of the model. A well-established approach to reduce the computational burden is K-fold crossvalidation ([14]). However, this method involves K model fits, so that it may still require a considerable amount of time, especially for large data sets. As a potential alternative, mixed predictive model checks are presented ([28]), where the model is fitted just once to the complete data set. In each step of the following crossvalidation, the estimated individual random effects and the concrete observation are ignored, and a forecast is generated based on the estimated fixed effects and the hyperparameters of the random effects. As the omitted observation influences the random effects only via their estimated covariance matrix, but not directly, the introduced conservatism is only moderate, and thus tolerable. This approach has been used before to select linear mixed models ([4]), and in different contexts ([34], [22]).

In order to conduct this predictive crossvalidation approach for generalized linear mixed models, each of the competing models is fitted only once to the

whole data set. After fitting the model, one observation y_{ij} from the data set is left out, and the predictive distribution for this observation is calculated based on the remaining observations $\boldsymbol{y}_{i,-j}$. Specifically, we apply a Bayesian iteratively weighted least squares (IWLS) algorithm for the calculation of the individual random effects $\hat{\boldsymbol{b}}_{i,-j}$ and their covariance matrix $\hat{\boldsymbol{Q}}_{i,-j}$. Note that this algorithm uses the originally estimated covariance of the random effects $\hat{\boldsymbol{Q}}$ as a priori information, for details see the following subsection. Apart from that, only the estimated fixed effects parameters $\hat{\boldsymbol{\beta}}$ are needed for the calculation of the predictive distribution, but not the initially estimated individual random effects parameters. To be more specific, the predicted expected value of the linear predictor is of the form

$$\mathrm{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T \hat{\boldsymbol{b}}_{i,-j} \tag{4}$$

with corresponding variance

$$\mathrm{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{z}_{ij}^T \hat{\boldsymbol{Q}}_{i,-j} \boldsymbol{z}_{ij}. \tag{5}$$

In order to stress the difference between this approach and a full leave-one-out crossvalidation, we show the predicted expected value and variance of the linear predictor obtained using full crossvalidation:

$$\mathrm{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_{-ij} + \boldsymbol{z}_{ij}^T \hat{\boldsymbol{b}}_{i,-j},$$

and

$$\mathrm{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{z}_{ij}^T \hat{\boldsymbol{Q}}_{-ij} \boldsymbol{z}_{ij}.$$

where all measurements except y_{ij} are used for the estimation of $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ and $\boldsymbol{Q}$.

The second step after having obtained (4) and (5) involves the calculation of the first two moments of the predictive distribution. This depends on the specific regression model and can either be done by approximation or numerical integration. In a final step, the proper scores described in Section 3 can be calculated. This procedure is then repeated for each observation y_{ij}.

The required steps for these calculation are discussed below for the case of binary logistic and log-linear Poisson regression with and without overdispersion.

A short discussion of its applicability in further generalized linear mixed models can be found in Section 7.

4.1. Bayesian iteratively weighted least squares algorithm

As discussed above, several steps are necessary to obtain the leave-one-out predictive distribution for the observation y_{ij}. First, estimates of the conditional expectation $\mathrm{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j})$ and covariance matrix $\mathrm{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j})$ are needed. To do this, the following algorithm is used ([39], [16]): Treat $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ as a given offset, and $\boldsymbol{z}_{ij}$ like "normal" covariates, so that a Bayesian estimation of the "regression coefficients" $\boldsymbol{b}_i$ can be performed. Combining the likelihood with the prior distribution $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \hat{\boldsymbol{Q}})$ leads to the approximate posterior distribution

$$\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j} \overset{a}{\sim} N(\boldsymbol{m}_{ij}, \boldsymbol{C}_{ij}),$$

whose parameters are obtained using the following Bayesian iteratively weighted least squares (IWLS) algorithm. Note that this algorithm is equivalent to a so-called penalized iteratively reweighted least squares (PIRLS) algorithm ([3]) and is also used in the `R` package `lme4` for the estimation of random effects in generalized linear mixed models ([2]). It works as follows: After choosing some starting values for $\boldsymbol{m}_{ij}^{(0)}$, for which we use the estimated random effects $\hat{\boldsymbol{b}}_i$ from the model fit, the estimates $\boldsymbol{m}_{ij}^{(k)}$ and $\boldsymbol{C}_{ij}^{(k)}$ in the kth iteration are

$$\boldsymbol{m}_{ij}^{(k)} = \boldsymbol{C}_{ij}^{(k)} \boldsymbol{z}_{i,-j} \boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) \tilde{\boldsymbol{y}}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}).$$

and

$$\boldsymbol{C}_{ij}^{(k)} = \{\hat{\boldsymbol{Q}}^{-1} + \boldsymbol{z}_{i,-j} \boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) \boldsymbol{z}_{i,-j}^T\}^{-1}$$

The "design matrix" $\boldsymbol{z}_{i,-j}$ of dimension $q \times (J-1)$ contains data from all time points of individual i except the time point of interest t_j. The elements of the response vector $\tilde{\boldsymbol{y}}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)})$ are the pseudo observations

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + \{y_{is} - \mu_{is}(\boldsymbol{m}_{ij}^{(k-1)})\} g'\{\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)})\} \qquad (6)$$

for $s \neq j$, where $\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = g^{-1}(\boldsymbol{x}_{is}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})$. The weights $W_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ are defined via ([16])

$$W_{is}^{-1}(\boldsymbol{m}_{ij}^{(k-1)}) = \kappa''(\theta_{is}(\boldsymbol{m}_{ij}^{(k-1)})) \{g'(\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)}))\}^2, \qquad (7)$$

so that the matrix containing all weights is $\boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) = \text{diag}\{W_{is}(\boldsymbol{m}_{ij}^{(k-1)})\}_{s \neq j}$.

These iterations are terminated as soon as

$$\max\left\{\frac{|\boldsymbol{m}_{ij}^{(k)} - \boldsymbol{m}_{ij}^{(k-1)}|}{|\boldsymbol{m}_{ij}^{(k-1)}|}\right\} < \epsilon,$$

with e.g. $\epsilon = 10^{-6}$, where $|\cdot|$ and max are taken over all components of $\boldsymbol{m}_{ij}^{(k)}$ and $\boldsymbol{m}_{ij}^{(k-1)}$. Thus, $\text{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{m}_{ij}$ and $\text{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{C}_{ij}$ are obtained. The specific formulae for (6) and (7) for different model types are shown in the following subsections.

4.2. Predictive crossvalidation for mixed logistic regression models

If applied to a mixed logistic regression model, formulae (6) and (7) have the form

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + \frac{y_{is} \cdot (1 + \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}))^2}{\exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})} - \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}) - 1$$

and

$$W_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \frac{\exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})}{(1 + \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}))^2},$$

leading to estimates $\text{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) = \hat{\boldsymbol{b}}_{i,-j} \approx \boldsymbol{m}_{ij}$ and $\text{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) = \hat{\boldsymbol{Q}}_{i,-j} \approx \boldsymbol{C}_{ij}$. The expected value and variance of η_{ij} have the form of equations (4) and (5). For further calculations, let $\text{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) =: \tau$ and $\text{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) =: \sigma^2$. In order to obtain the predictive probability $P\{y_{ij} = 1 \,|\, \boldsymbol{y}_{i,-j}\}$, the mixed logistic regression model is rewritten as a latent variable model. Expression (1) corresponds to

$$\omega_{ij} = \eta_{ij} + \epsilon_{ij},$$

where $y_{ij} = 1$ if $\omega_{ij} \geq 0$, $y_{ij} = 0$ if $\omega_{ij} < 0$ and ϵ_{ij} follows a standard logistic distribution. This can be approximated by a normal distribution ([43]), so that

$$\epsilon_{ij} \overset{a}{\sim} N(0, c)$$

with $c = (15/16)^2 \cdot \pi^2/3$. Thus,

$$\omega_{ij} \overset{a}{\sim} N(\tau, \sigma^2 + c),$$

so that $P\{y_{ij} = 1 \,|\, \boldsymbol{y}_{i,-j}\} = \int_0^\infty N(x \,|\, \tau, \sigma^2 + c)dx$ can be calculated based on the distribution function of the normal distribution and subsequently used for the calculation of the BS and the LS.

4.3. Predictive crossvalidation for mixed Poisson regression models

In the case of a log-linear Poisson regression model, the pseudo observations $\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ for $s \neq j$ are

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + y_{is} \exp(-\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} - \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}) - 1, \tag{8}$$

and the weights $W_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ are

$$W_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}). \tag{9}$$

The calculation of the first two moments of the predictive distribution is straightforward. Expected value τ and variance σ^2 of η_{ij} are again obtained as in (4) and (5). As η_{ij} is (approximately) normally distributed, $\exp(\eta_{ij}) = \lambda_{ij}$ is log-normally distributed with

$$\mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big) \tag{10}$$

and

$$\mathrm{Var}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \{\exp(\sigma^2) - 1\} \exp(2\tau + \sigma^2). \tag{11}$$

In a final step, the predictive expectation and variance of the observation y_{ij} are

$$\mathrm{E}(y_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big)$$

and

$$\begin{aligned}
\mathrm{Var}(y_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \mathrm{E}\{\mathrm{Var}(y_{ij} \,|\, \eta_{ij}, \boldsymbol{y}_{i,-j})\} + \mathrm{Var}\{\mathrm{E}(y_{ij} \,|\, \eta_{ij}, \boldsymbol{y}_{i,-j})\} \\
&= \mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) + \mathrm{Var}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) \\
&= \exp\Big(\tau + \frac{1}{2}\sigma^2\Big) + \{\exp(\sigma^2) - 1\} \exp(2\tau + \sigma^2).
\end{aligned}$$

These two values allow the calculation of the DSS (3) for each observation from the data set and its respective prediction, and subsequently of the mean DSS.

To obtain the LS, however, the predictive expectation and variance are not sufficient, because the density of the predictive distribution has to be known. This problem can be solved using two distinct approaches: The first possibility is an approximation of the log-normal distribution of $\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}$ via the gamma distribution, which is performed by matching the first two moments of these two distributions:

The two parameters of the gamma distribution can be chosen in such a way that its expected value and variance equal (10) and (11), respectively. This is the case for a gamma distribution $G(\alpha, \phi)$ with density

$$f(\lambda_{ij}) = \frac{1}{\Gamma(\alpha)} \Big(\frac{\alpha}{\phi}\Big)^{\alpha} \lambda_{ij}^{\alpha-1} \exp\Big(-\frac{\lambda_{ij}\alpha}{\phi}\Big)$$

and parameters

$$\alpha = \frac{\mathrm{E}(\lambda_{ij})^2}{\mathrm{Var}(\lambda_{ij})} = \frac{1}{\exp(\sigma^2) - 1}$$

and

$$\phi = \mathrm{E}(\lambda_{ij}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big).$$

Note that this approach works only if $\alpha > 1$, because the respective gamma distribution must have a mode larger than 0.

With λ_{ij} being approximately gamma distributed, the marginal distribution of y_{ij} follows a negative binomial distribution ([40, p. 35]) with density

$$\begin{aligned} f(y_{ij}) &= \int_0^{\infty} f(y_{ij} \,|\, \lambda_{ij}) f(\lambda_{ij}) d\lambda_{ij} \\ &= \frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha)\Gamma(y_{ij} + 1)} \Big(\frac{\alpha}{\phi + \alpha}\Big)^{\alpha} \Big(\frac{\phi}{\phi + \alpha}\Big)^{y_{ij}}. \end{aligned}$$

Evaluating this (log) density with mean parameter ϕ and size parameter α at the actual observation yields the desired LS. To ensure that the used approximation steps work reasonably well, we recommend comparing the resulting mean LS with the mean DSS.

Alternatively, one can simply use numerical integration, which should be reasonably quick if the data set is not too large. In that case, the density of the predictive distribution at the actual observation y_{obs} is obtained by the integral

$$f(y_{\text{obs}}) = \int_0^\infty f(y_{\text{obs}} \,|\, \lambda_{ij}) f(\lambda_{ij}) d\lambda_{ij}, \tag{12}$$

where λ_{ij} follows a log-normal distribution with parameters (10) and (11).

4.4. Predictive crossvalidation for mixed Poisson regression models with overdispersion

The predictive crossvalidation procedure with overdispersion works almost as in the ordinary Poisson case (without overdispersion), with some minor changes: Let

$$\boldsymbol{\eta}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{r}_i \boldsymbol{d}_i$$

be the model for all observations of individual i, where

$$\boldsymbol{d}_i = \begin{pmatrix} \boldsymbol{b}_i \\ \boldsymbol{a}_i \end{pmatrix}$$

is the vector containing all random effects of individual i, the design matrix of the fixed effects is

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \vdots \\ \boldsymbol{x}_{iJ}^T \end{pmatrix}$$

and the design matrix for the random effects has the form

$$\boldsymbol{r}_i = \begin{pmatrix} \boldsymbol{z}_{i1}^T & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \boldsymbol{z}_{iJ}^T & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

For the estimation of the conditional expected value of $\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j}$, we use the prior distribution $\boldsymbol{d}_{ij} \sim N(\boldsymbol{0}, \text{diag}(\hat{\boldsymbol{Q}}, \hat{\xi}^2, \ldots, \hat{\xi}^2))$, and $\boldsymbol{r}_{i,-j}$ is $\boldsymbol{r}_i$ without the jth row and has to be used instead of $\boldsymbol{z}_{i,-j}$ in formulae (8) and (9). Apart from that, all remaining formulae from the IWLS algorithm stay the same. Note that the calculation of the expected value and variance of η_{ij} as in formulae (4) and

(5) are now

$$\begin{aligned}\mathrm{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{r}_{ij}^T \,\mathrm{E}(\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j}) \\ &= \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T \,\mathrm{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) + 0\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \boldsymbol{r}_{ij}^T \,\mathrm{Cov}(\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j})\boldsymbol{r}_{ij} \\ &= \boldsymbol{z}_{ij}^T \,\mathrm{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j})\boldsymbol{z}_{ij} + \hat{\xi}^2,\end{aligned}$$

meaning that the expected value and variance of $a_{ij} \,|\, \boldsymbol{y}_{i,-j}$ are the moments of its prior distribution, and only the predictive expectation and variance of $\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}$ need to be calculated.

5. Simulation study

To evaluate the properties of our proposed crossvalidation method for model selection, we performed a simulation analysis with both log-linear Poisson and binary logistic regression models. In both cases we looked at different numbers of individuals ($I = 10$ or 50) and numbers of measurements per individual ($J = 10$, 20 or 40). Combining these numbers lead to five different settings, where the combination of 50 individuals with 40 measurements each was omitted for time reasons. For each setting, 100 data sets were generated and evaluated.

Binary data for logistic regression were generated from a model with three fixed effects and one random intercept. The fixed covariates were time (1 to number of measurements per individual, standardized and centered around 0) and two binary (time-independent) covariates x_2 and x_3 which are independently Bernoulli(0.5) distributed and also centered around 0; the intercept was set to 0. The corresponding coefficients were $\beta_{\text{time}} = \beta_2 = \beta_3 = 1$, and the random intercept b_i was generated from a $N(0, 0.25)$ distribution. Count data for Poisson regression were generated from a similar model, the only differences are that x_2 and x_3 were 0 or 1, i.e. not centered, and an intercept of 2 was added to the linear predictor.

In both situations we fitted three different models to each data set: The "true" model with three fixed effects and a random intercept (denoted by "(3,1)", representing the number of fixed and random effects), a model without x_3 (denoted by "(2,1)") and a model with all three fixed effects, the random intercept and an additional random slope over time (denoted by "(3,2)"). For each model we calculated the mean LS and the mean BS or DSS, respectively. For comparison, the conditional AIC (cAIC, [13]) was calculated. Note that in some cases the estimated covariance matrix of the random effects was singular and thus not invertible. In these cases a score of $-\infty$ was attributed. For a few of the generated data sets, none of the three competing models could be fitted due to numerical problems of the `lme4` function. If that was the case, we generated a new data set, until 100 valid comparisons per setting were obtained.

Tables 1 and 2 show the results of this simulation study for binary logistic and Poisson regression. The true model is in both cases the one in the middle column. For logistic regression, we can see that for a small number of individuals I, our method performs better than the cAIC, which often prefers simpler models. This behaviour seems to be independent of the number of observations per person. If I increases, all three criteria show a similar performance: they clearly reject the model (2,1), but choose the wrong model with two random effects in almost two thirds of the cases.

The results of the Poisson regression are comparable. The mean LS selects the correct model more often than cAIC, with the mean DSS being in between if I is small and worse than the other two if I increases. In summary, the cAIC tends to select simpler models. The mean LS performs better than cAIC if the number of individuals is small, whereas for a larger I, both criteria come to similar results. The BS is comparable to the LS, however, the DSS is slightly worse.

In addition to the mean scores, we also looked at the estimated coefficients of the fixed effects. Ideally, a good method chooses models that provide a satisfying amount of coverage. To evaluate the abilities of our proposed method and avoid misleading results due to Monte Carlo error, we calculated z-values

Table 1: Logistic regression: Percentage of chosen models for simulated data

	(2,1)	(3,1)	(3,2)
10 individuals, 5 observations:			
LS	14	43	43
BS	15	43	42
cAIC	42	22	36
10 individuals, 10 observations:			
LS	9	45	46
BS	9	48	43
cAIC	27	27	46
10 individuals, 40 observations:			
LS	3	50	47
BS	4	50	46
cAIC	25	34	41
50 individuals, 5 observations:			
LS	1	36	63
BS	1	41	58
cAIC	1	37	62
50 individuals, 10 observations:			
LS	1	35	64
BS	1	34	65
cAIC	1	35	64

Table 2: Poisson regression: Percentage of chosen models for simulated data

	(2,1)	(3,1)	(3,2)
10 individuals, 10 observations:			
LS	7	62	31
DSS	16	56	28
cAIC	20	50	30
10 individuals, 20 observations:			
LS	4	61	35
DSS	16	50	34
cAIC	18	45	37
10 individuals, 40 observations:			
LS	7	58	35
DSS	15	51	34
cAIC	24	42	34
50 individuals, 10 observations:			
LS	2	47	51
DSS	9	42	49
cAIC	2	46	52
50 individuals, 20 observations:			
LS	1	52	47
DSS	9	49	42
cAIC	1	51	48

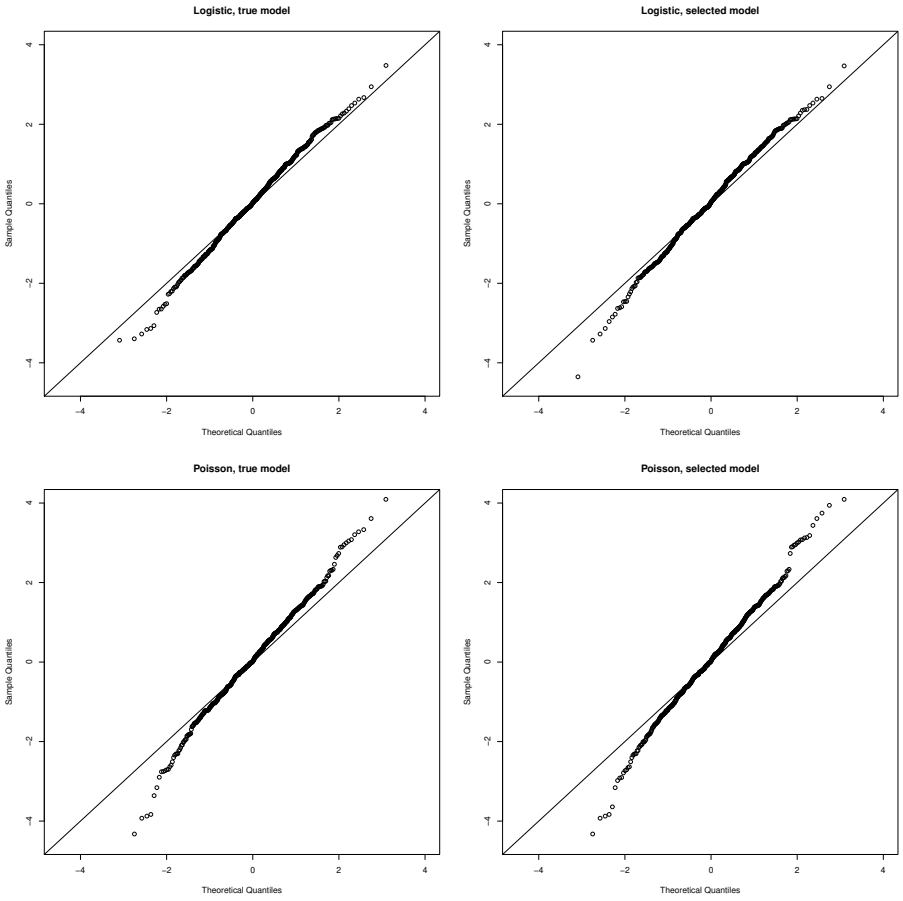

Figure 1: Q-Q plot of the z-values of the coefficients β_2 of the logistic (upper row) and Poisson regression models (lower row); left column: true model; right column: selected model

$(\hat{\beta}_2 - \beta_2)/\text{se}(\hat{\beta}_2)$ from the models fitted to the simulated data sets, here $\text{se}(\hat{\beta}_2)$ denotes the standard error of the estimate $\hat{\beta}_2$ of $\beta_2 = 1$. These z-values should be approximately standard normally distributed.

Figure 1 shows the corresponding Q-Q plots of the z-values from both the true models (left) and the models selected by our method using the mean LS (right). In the upper row you can find the plots from the logistic regression models, whereas in the lower row are the plots which correspond to the Poisson regression models. We see that in both cases, the Q-Q plots from the true and the selected models look very much alike, so that we can conclude that our method works well in estimating the fixed effects even if we take the model selection process into account.

6. Application

We illustrate the methods presented above using two well-known data sets from the literature. Although both data sets have been analysed many times, no formal selection of the random effects was possible so far due to lack of a suitable criterion for this task. It was therefore also not possible to choose fixed and random effects at the same time. The data and `R` functions used for the following analyses can be found as supplementary material, along with explanations of the functions.

6.1. Case study I: Xerophthalmia and respiratory disease in Indonesian children

The first data set is presented by Diggle et al. [12, p. 4] and contains binary data on infections of the respiratory tract and xerophthalmia (dryness of the eye due to vitamin A deficiency) in Indonesian children, along with additional information on the children's age and height. Up to 6 measurements per child were collected in quarterly visits. 22 of the 275 patients were removed because they contributed only one measurement, so that the remaining data set consists of 1178 measurements of 253 children with 105 events of respiratory infection.

The variables sex, height in relation to age (percentage obtained from the United States National Center for health statistics), and presence of xerophthalmia are included in all models. As suggested by Diggle et al. [12, p. 156

and 182], a model M1 with age and age squared was compared to a model M2 with follow-up time and follow-up time squared instead. In the latter case, age at baseline and age squared at baseline are included as additional covariates. In addition, both models were fitted with two covariates representing an annual sine and cosine (denoted by "with season"). All models either include just a random intercept (RI) or, alternatively, an additional random slope (RIS) which depends on age (M1) or time (M2).

The mean BS and LS of these eight models are given in Table 3. Both scores clearly prefer models including a seasonal sine and cosine to the ones without that seasonal component, and the models with the lowest BS or, respectively, LS both use follow-up time instead of age. The BS chooses a model with just random intercept (BS: −0.07537), whereas the LS selects the model with additional random slope (LS: −0.27376).

Table 3: Respiratory infection: Mean BS and LS for the eight different models

	BS	LS
M1 (age), RI	−0.07732	−0.28124
M1 (age), RIS	−0.07695	−0.27917
M1 with season, RI	−0.07593	−0.27622
M1 with season, RIS	−0.07557	−0.27435
M2 (time), RI	−0.07644	−0.27678
M2 (time), RIS	−0.07665	−0.27680
M2 with season, RI	−0.07537	−0.27386
M2 with season, RIS	−0.07567	−0.27376

6.2. Case study II: Seizure counts

The second example analyses a frequently used data set ([37], [12, p. 10]), with counts of epileptic seizures as outcome. In this randomized crossover study patients were treated against partial epileptic seizures with either progabide (an anti-epileptic drug) or placebo and followed over four subsequent clinic visits. At each visit, they reported the number of epileptic seizures during the last two weeks. In our analyses, only the clinic visits before crossing over to the alternative treatment are analysed. Note that one patient was left out due to very unusual measurements, as suggested by Diggle et al. [12, p. 164].

The data set contains information of 58 patients with four clinic visits each. The covariates used in all models are baseline seizure rate, treatment as well as a baseline-treatment interaction term, and the logarithm of age, completed by either the respective visit number or its square root serving as time variable. The following models are compared: One model with just a random intercept and a second model comprising an additional random slope, either using the visit number or its square root in both the fixed and the random effects part. In a second step, we add an indicator for the fourth visit to each of these models in order to account for markedly low counts at the last visit of each patient.

The first and second column of Table 4 display the mean crossvalidated DSS and LS of the eight competing models. We used the approximation of the log-normal distribution by a gamma distribution as explained in Section 4.3. For comparison, we also used numerical integration as in formula (12) and calculated the same integral with the approximate gamma distribution for λ_{ij}. All three methods lead to practically the same results. Note that in this example, the DSS and the LS put the models in exactly the same order. For both versions of the time variable, the model with random intercept and slope is preferred, and both scores show a clear preference for models including the indicator of the fourth visit. Finally, using the square root of time instead of the original time variable leads to an additional improvement, in particular concerning the models with indicator of the fourth visit. All this indicates that the model best suited for prediction is the model with random intercept and slope, based on the square root of the time variable and including a variable indicating the fourth visit, having a mean DSS of -1.9646 and a mean LS of -2.7536.

The third and fourth column of Table 4 show the mean scores for models with random intercept that include a random effect for each observation in order to incorporate overdispersion (denoted by "OD"). Fitting models with random intercept, slope and the random effect for each observation leads to overfitting and causes the variance of the random slope to be very small and the correlation between random slope and intercept to be 1. From this we can conclude that including an additional random slope does not ameliorate the overdispersion

model.

Concerning the models with random intercept only, we can see that accounting for overdispersion leads to a remarkable improvement in both mean scores. In contrast, the differences between the four models are small, showing that all four models are equally useful for making predictions. Both scores again select models that include the indicator for the fourth visit, and the LS chooses the model with visit as time variable (LS: −2.5617), whereas the DSS slightly prefers the model with the square root of visit number (DSS: −1.766).

Table 4: Epileptic seizures: Mean DSS and LS for the eight different models with and without overdispersion (OD). "−" indicates that scores could not be calculated due to a singular covariance matrix.

	DSS	LS	DSS, OD	LS, OD
visit, RI	−2.0453	−2.7946	−1.7676	−2.5618
visit, RIS	−2.0165	−2.7841	−	−
sqrt(visit), RI	−2.0487	−2.7962	−1.7696	−2.5625
sqrt(visit), RIS	−2.0012	−2.7706	−	−
visit, RI, visit 4	−2.0153	−2.7867	−1.7669	−2.5617
visit, RIS, visit 4	−1.9846	−2.7705	−	−
sqrt(visit), RI, visit 4	−2.0152	−2.7867	−1.7660	−2.5618
sqrt(visit), RIS, visit 4	−1.9646	−2.7536	−	−

6.3. Comparison with other methods

We compare the results of our proposed method with the ones obtained using two other suggestions. The first alternative method ([13]) involves an asymptotic version of the cAIC. The second method we compare our results to is the corrected conditional AIC (ccAIC, [42]), which can also be applied if the covariance matrix of the random effects is unknown. The authors kindly provided us with their Matlab programs which we translated into `R` functions and extended to the case of binary logistic regression. Unfortunately, these programs are so far only useable for models with just a random intercept and no random slope, for which reason the ccAIC could not be calculated for all candidate models in our applications. These two alternative methods can so far not be used for a Poisson model including an additional random effect to cover

overdispersion. Note that both criteria seem to be very similar, which is also confirmed by the results in our applications.

Table 5: Logistic regression: Comparison with other methods; (c)cAIC values transformed by $-\frac{1}{2n}$

	LS	cAIC Donohue	ccAIC Yu
M1 (age), RI	−0.28124	−0.28254	−0.28254
M1 (age), RIS	−0.27917	−0.27868	–
M1 with season, RI	−0.27622	−0.27903	−0.27903
M1 with season, RIS	−0.27435	−0.27504	–
M2 (time), RI	−0.27678	−0.27954	−0.27955
M2 (time), RIS	−0.27680	−0.27802	–
M2 with season, RI	−0.27386	−0.27786	−0.27787
M2 with season, RIS	−0.27376	−0.27585	–

Table 6: Poisson regression: Comparison with other methods; (c)cAIC values transformed by $-\frac{1}{2n}$

	LS	cAIC Donohue	ccAIC Yu
visit, RI	−2.7946	−2.6686	−2.6687
visit, RIS	−2.7841	−2.6090	–
sqrt(visit), RI	−2.7962	−2.6698	−2.6699
sqrt(visit), RIS	−2.7706	−2.5992	–
visit, RI, visit 4	−2.7867	−2.6672	−2.6672
visit, RIS, visit 4	−2.7705	−2.6023	–
sqrt(visit), RI, visit 4	−2.7867	−2.6672	−2.6672
sqrt(visit), RIS, visit 4	−2.7536	−2.5905	–

Tables 5 and 6 again show the mean LS obtained with our proposed procedure, along with the cAIC ([13], denoted by "Donohue") and the ccAIC ([42], denoted by "Yu", only for models with random intercept). To ease the comparisons, we put the cAIC and ccAIC values on the same scale as the mean LS by dividing them by $-2\sum J_i$, as has been done before ([4]).

Concerning the logistic regression model in Table 5, the results from the three different procedures differ only in the third decimal place. The models with just a random intercept are arranged in the same order by all three methods. If all models are taken into account (i.e. a random slope as well), the best three models and the worst model are clearly identified by our method and the cAIC, however, the overall order is slightly different and another model is chosen to

be the best.

If we have a look at Table 6, we can see that the differences between our proposed method and the other two methods are larger than for logistic regression, but the results are still of similar magnitude. The ordering of the competing models is not identical but similar, and especially the decision which model is the best or worst is the same using all three methods. In order to find an explanation for the differences between mean LS and cAIC and ccAIC, we conducted some additional simulation studies. We found that if the true distribution of the data is Poisson and a Poisson model is fitted, there are only very small differences between the different methods. If, however, the data are overdispersed - as it is the case in our application - and come from a negative binomial distribution, the differences increase along with the amount of overdispersion.

To illustrate the comparison between the transformed mean LS and the cAIC, Figure 2 shows the respective values from both methods in both applications. We do not show the ccAIC, as there are no visible differences between that and cAIC. In both cases, the models with just random intercept perform worse than models with additional random slope. The order of the models with random intercept is the same, but small differences occur when a random slope is included. Summing up, our proposed crossvalidation approach leads to results that are comparable to the other two possible methods, but not exactly the same, especially for count data with overdispersion.

6.4. Comparison with full crossvalidation

We additionally compare the results obtained with our approximate crossvalidation approach with those from a full crossvalidation. The two plots in Figure 3 show the mean DSS and LS values from the applications of log-linear Poisson regression models. We can see that the scores from the approximate crossvalidation are better than from a full crossvalidation, which is the expected behaviour. All points are relatively close to the diagonal, so we conclude that our approximate crossvalidation procedure is reliable enough in the case of log-linear Poisson regression.

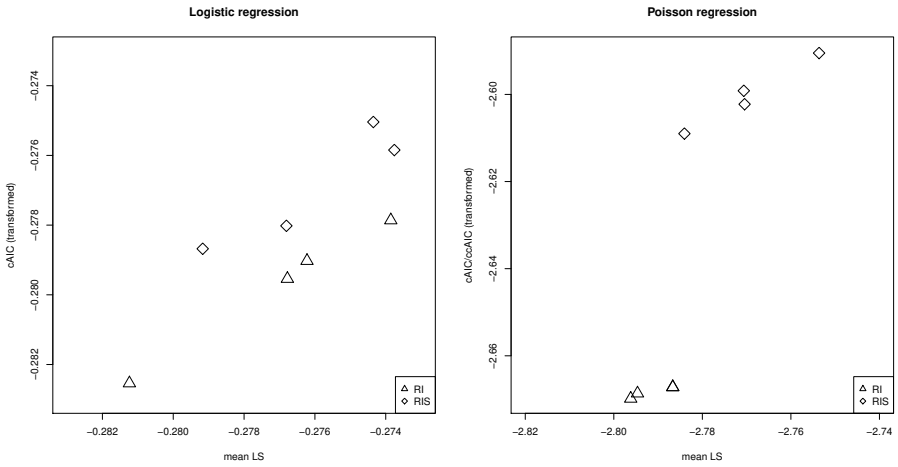

Figure 2: Comparison of transformed cAIC with mean LS for models with random intercept (RI) and models with additional random slope (RIS). Note that the values of two models with random intercept in the right plot are equal, for which reason only three RI models are visible.

Unfortunately, we cannot show a similar comparison in the logistic regression case study. The reason for this is a problem that often occurs during the fitting process in the `lme4` function: The covariance matrix of the random effects is often estimated to be not invertible in models with random intercept and slope. In these cases, the predictive distribution and the according scores can not be calculated, and a value of $-\infty$ must be implicitly assigned to the respective scores. This phenomenon occured very often during the full crossvalidations, so that there are between 184 and 420 scores of in total 1178 observations missing. Consequently, the resulting mean scores are not reliable, which is a clear disadvantage of the full in comparison to the approximate crossvalidation.

Apart from the fact that results from the approximate crossvalidation approach are more reliable in cases where a full crossvalidation would yield many singular estimated covariance matrices, the time is also an important point in favor of our suggested approach. In the Poisson regression example, our approach needed 1-2 seconds, whereas the full crossvalidation lasted 5.5 minutes, a considerable difference for a relatively small data set of 232 observations. With

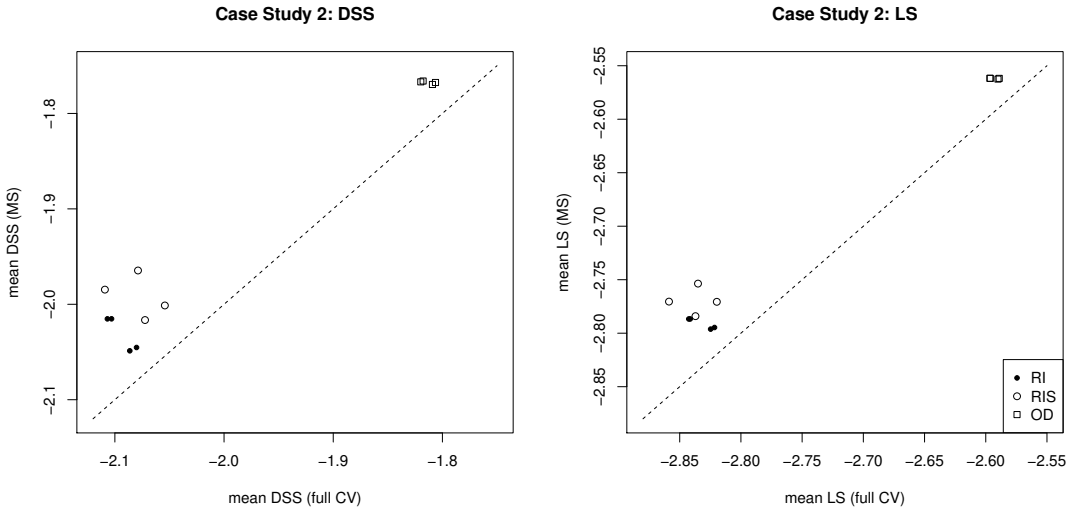

Figure 3: Comparison of our approach (MS) with a full crossvalidation (full CV)

a larger data set the time difference would certainly be even more pronounced.

7. Discussion

This paper has presented a novel predictive crossvalidation approach to model selection in generalized linear mixed models. The crossvalidated LS and BS or DSS form a useful means for selecting both fixed and random effects. As the model has to be fitted only once, this approach is much less time-consuming than a true leave-one-out crossvalidation.

We have demonstrated the calculation of mean proper scoring rules with our crossvalidation approach for the two most common generalized linear mixed models, i.e. logistic regression and Poisson regression (with or without overdispersion), but the approach is applicable more widely, often using procedures that are very similar to the ones used above. For example, overdispersed binomial data can be analysed by adding an additional random effect for each observation to the linear predictor, and the predictive distribution is then obtained by using the formulation as latent variable model from Section 4.2. The first two

moments of the predictive distribution in log-linear gamma regression models can be calculated analogously to Poisson regression models, and numerical integration as in (12) can be used for the calculation of the LS. The derivation of the predictive distribution or its moments could be slightly more complicated in some other, more rarely used generalized linear mixed models, but it should be feasible in most cases.

The application of the IWLS algorithm in the Poisson case with overdispersion can cause problems in data sets with a large number of observations, because inversions of large matrices are needed. However, in our predictive crossvalidation approach it is only applied for one individual in the data set at a time, so this should not be problematic. Note that an alternative algorithm based on building blocks of correlated parameters is provided ([16]), which can be used for larger data sets if needed.

An alternative approach to modelling overdispersion would be to replace the Poisson with a negative binomial model. Unfortunately, adapting the Bayesian IWLS algorithm to this setting would be difficult. Moreover, it is not clear how the overdispersion parameter of the negative binomial distribution could be estimated in a mixed model setting. In contrast, our proposed method makes use of normal random effects so that the model can always be fitted using existing software.

All proper scoring rules used in this paper are calculated for one observation and the respective univariate predictive distribution only. Multivariate versions of several proper scores that can be calculated for a whole set of observations and their multivariate predictive distribution are presented ([19]). This could theoretically be used to conduct K-fold crossvalidation as mentioned at the beginning of Section 4, which would be much more time-consuming than our approach, but might be desired in certain situations.

Concerning the calculation of the predictive density for obtaining the LS for a mixed Poisson model, the approximation of the log-normal distribution via a gamma distribution can be realized in different ways. The matching of moments which we have applied could be problematic for certain forms of the

respective distributions. As a possible alternative, we have tried minimizing the Kullback-Leibler distance between the two distributions, but this is much more time-consuming and often not applicable due to numerical problems. For this reason, it is advisable to calculate both the LS and the DSS in the case of a mixed Poisson model and see if the results are comparable.

In comparison with other possible approaches to model choice in generalized linear mixed models, our method has two decisive advantages: First, it can take into account overdispersion, which proves very useful in routine applications. Second, other approaches involve the multiplication and inversion of large matrices. This is not a problem in small data sets, but as soon as the data set gets large, the necessary calculations take considerable time or may not be possible at all.

Acknowledgements

We thank Sarah Haile for proofreading this article. We also thank two anonymous reviewers and the associate editor for valuable comments that helped to improve an earlier version of this article.

References

[1] Akaike, H., 1973. Information theory and extension of the maximum likelihood principle, in: Petrov, B.N., Csaki, F. (Eds.), International Symposium on Information Theory, Budapest: Akademia Kiado. pp. 267–281.

[2] Bates, D., 2012. Linear mixed model implementation in `lme4`. http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf.

[3] Bates, D., DebRoy, S., 2004. Linear mixed models and penalized least squares. Journal of Multivariate Analysis 91, 1–17.

[4] Braun, J., Held, L., Ledergerber, B., 2012. Predictive cross-validation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study. Biometrics 68, 53–61.

[5] Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3.

[6] Cai, B., Dunson, D., 2006. Bayesian covariance selection in generalized linear mixed models. Biometrics 62, 446–457.

[7] Claeskens, G., Hjort, N., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge, Great Britain.

[8] Collett, D., 2003. Modelling Binary Data. Chapman & Hall/CRC. 2nd edition.

[9] Crainiceanu, C.M., Ruppert, D., 2004. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society, Ser. B 66, 165–185.

[10] Czado, C., Gneiting, T., Held, L., 2009. Predictive model assessment for count data. Biometrics 65, 1254–1261.

[11] Dawid, A.P., Sebastiani, P., 1999. Coherent dispersion criteria for optimal experimental design. The Annals of Statistics 27, 65–81.

[12] Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L., 2002. Analysis of Longitudinal Data. Oxford University Press. 2nd edition.

[13] Donohue, M., Overholser, R., Xu, R., Vaida, F., 2011. Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika 98, 685–700.

[14] Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

[15] Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer. second edition.

[16] Gamerman, D., 1997. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing 7, 57–68. 10.1023/A:1018509429360.

[17] Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society, Ser. B 69, 243–268.

[18] Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102, 359–378.

[19] Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17, 211–235.

[20] Greven, S., Kneib, T., 2010. On the behavior of marginal and conditional AIC in linear mixed models. Biometrika 97, 773–789.

[21] Held, L., Rufibach, K., Balabdaoui, F., 2010a. A score regression approach to assess calibration of continuous probabilistic predictions. Biometrics 66, 1295–1305.

[22] Held, L., Schrödle, B., Rue, H., 2010b. Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA., in: Kneib, T., Tutz, G. (Eds.), Statistical Modelling and Regression Structures - Festschrift in Honour of Ludwig Fahrmeir. Physica-Verlag, Heidelberg, Germany, pp. 91–110.

[23] Hodges, J.S., Sargent, D.J., 2001. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika 88, 367–379.

[24] Jiang, J., Rao, J., Gu, Z., Nguyen, T., 2008. Fence methods for mixed model selection. The Annals of Statistics 36, 1669–1692.

[25] Lian, H., 2012. A note on conditional Akaike information for Poisson regression with random effects. Electronic Journal of Statistics 6, 1–9.

[26] Liang, H., Wu, H., Zou, G., 2008. A note on conditional AIC for linear mixed-effects models. Biometrika 95, 773–778.

[27] Lu, H., Hodges, J.S., Carlin, B.P., 2007. Measuring the complexity of generalized linear hierarchical models. The Canadian Journal of Statistics 35, 69–87.

[28] Marshall, E.C., Spiegelhalter, D.J., 2003. Approximate cross-validatory predictive checks in disease mapping models. Statistics in Medicine 22, 1649–1660.

[29] Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer.

[30] Nguyen, T., Jiang, J., 2012. Restricted fence method for covariate selection in longitudinal data analysis. Biostatistics 13, 303–314.

[31] Pauler, D., 1998. The Schwarz criterion and related methods for normal linear models. Biometrika 85, 13–27.

[32] Pauler, D., Wakefield, J., Kass, R., 1999. Bayes factors and approximations for variance component models. Journal of the American Statistical Association 94, 1242–1253.

[33] Pawitan, Y., 2001. In All Likelihood - Statistical Modelling and Inference Using Likelihood. Oxford University Press, Oxford, Great Britain.

[34] Riebler, A., Held, L., 2010. The analysis of heterogeneous time trends in multivariate age-period-cohort models. Biostatistics 11, 57–69.

[35] Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

[36] Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, Ser. B 39, 44–47.

[37] Thall, P.F., Vail, S.C., 1990. Some covariance models for longitudinal count data with overdispersion. Biometrics 46, 657–671.

[38] Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed-effects models. Biometrika 92, 351–370.

[39] West, M., 1985. Generalized linear models: Scale parameters, outlier accommodation, scale parameters and prior distributions, in: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics 2, Elsevier Science Publishers B. V., North Holland. pp. 531–558.

[40] Winkelmann, R., 2008. Econometric Analysis of Count Data. Springer.

[41] Winkler, R., 1996. Scoring rules and the evaluation of probabilities. Test 5, 1–60.

[42] Yu, D., Yau, K.K.W., 2012. Conditional Akaike information criterion for generalized linear mixed models. Computational Statistics and Data Analysis 56, 629–644.

[43] Zeger, S.L., Liang, K.Y., Albert, P.S., 1988. Models for longitudinal data: A generalized estimating equation approach. Biometrics 44, 1049–1060.